

smg



UNIVERSITY OF
SURREY

Computational morphology needs better lexical data

Sacha Beniamine Webinaire ILFC, 16th of April 2025

Inflection

8

étudier

1^{er}

groupe

verbes en -ier

INDICATIF

• PRÉSENT

j' **étudie**
tu **étudies**
il/elle **étudie**
nous **étudions**
vous **étudiez**
ils/elles **étudient**

• IMPARFAIT

j' **étudiais**
tu **étudiais**
il/elle **étudiait**
nous **étudiions**
vous **étudiez**
ils/elles **étudiaient**

• PASSÉ SIMPLE

j' **étudiai**
tu **étudias**
il/elle **étudia**
nous **étudiâmes**
vous **étudiâtes**
ils/elles **étudièrent**

• FUTUR SIMPLE

j' **étudierai**
tu **étudieras**
il/elle **étudiera**
nous **étudierons**
vous **étudierez**

• PASSÉ COMPOSÉ

j' **ai étudié**
tu **as étudié**
il/elle **a étudié**
nous **avons étudié**
vous **avez étudié**
ils/elles **ont étudié**

• PLUS-QUE-PARFAIT

j' **avais étudié**
tu **avais étudié**
il/elle **avait étudié**
nous **avions étudié**
vous **aviez étudié**
ils/elles **avaient étudié**

• PASSÉ ANTERIEUR

j' **eus étudié**
tu **eus étudié**
il/elle **eut étudié**
nous **eûmes étudié**
vous **eûtes étudié**
ils/elles **eurent étudié**

• FUTUR ANTERIEUR

j' **aurai étudié**
tu **auras étudié**
il/elle **aura étudié**
nous **aurons étudié**
vous **aurez étudié**
ils/elles **auront étudié**

8

apprécier · copier · crier · envier ·
modifier · oublier · skier · trier...

SUBJONCTIF

• PRÉSENT

que j' **étudie**
que tu **étudies**
qu'il/elle **étudie**
que nous **étudions**
que vous **étudiez**
qu'ils/elles **étudient**

• IMPARFAIT

que j' **étudiassé**
que tu **étudiasses**
qu'il/elle **étudiât**
que nous **étudiassions**
que vous **étudiassiez**
qu'ils/elles **étudiassent**

• PASSÉ

que j' **aie étudié**
que tu **aies étudié**
qu'il/elle **ait étudié**
que nous **ayons étudié**
que vous **ayez étudié**
qu'ils/elles **aient étudié**

• PLUS-QUE-PARFAIT

que j' **eusse étudié**
que tu **eusses étudié**
qu'il/elle **eût étudié**
que nous **eussions étudié**
que vous **eussiez étudié**
qu'ils/elles **eussent étudié**

IMPÉRATIF

• PRÉSENT

étudie
étudions
étudiez

• PASSÉ

aie étudié
ayons étudié
ayez étudié

INFINITIF

• PRÉSENT

étudier

• PASSÉ

avoir étudié

PARTICIPE

• PRÉSENT

étudiant

• PASSÉ

ayant étudié
étudié(e, s, es)

Inflection

8

étudier

1^{er} groupe verbes en -ier

INDICATIF

● PRÉSENT

j' **étudie**
tu **étudies**
il/elle **étudie**
nous **étudions**
vous **étudiez**
ils/elles **étudient**

● IMPARFAIT

j' **étudiais**
tu **étudiais**
il/elle **étudiait**
nous **étudiions**
vous **étudiez**
ils/elles **étudiaient**

● PASSÉ SIMPLE

j' **étudiai**
tu **étudias**
il/elle **étudia**
nous **étudiâmes**
vous **étudiâtes**
ils/elles **étudièrent**

● FUTUR SIMPLE

j' **étudierai**
tu **étudieras**
il/elle **étudiera**
nous **étudierons**
vous **étudierez**

● PASSÉ COMPOSÉ

j' **ai étudié**
tu **as étudié**
il/elle **a étudié**
nous **avons étudié**
vous **avez étudié**
ils/elles **ont étudié**

● PLUS-QUE-PARFAIT

j' **avais étudié**
tu **avais étudié**
il/elle **avait étudié**
nous **avions étudié**
vous **aviez étudié**
ils/elles **avaient étudié**

● PASSÉ ANTÉRIEUR

j' **eus étudié**
tu **eus étudié**
il/elle **eut étudié**
nous **eûmes étudié**
vous **eûtes étudié**
ils/elles **eurent étudié**

● FUTUR ANTÉRIEUR

j' **aurai étudié**
tu **auras étudié**
il/elle **aura étudié**
nous **aurons étudié**
vous **aurez étudié**
ils/elles **auront étudié**

1^{re} DÉCLINAISON DES NOMS

(FÉMININS — quelques masculins)

GEN. SG. : -A E

12. Type : **rosa, ae, f.** : la rose, une rose.

		Terminaisons	
SINGULIER -	NOMINATIF	rosa	-a
	VOCATIF	rosa	-a
	ACCUSATIF	rosam	-am
	GÉNITIF	rosae	-ae
	DATIF	rosae	-ae
	ABLATIF	rosā	-ā
PLURIEL -	NOMINATIF	rosae	-ae
	VOCATIF	rosae	-ae
	ACCUSATIF	rosas	-as
	GÉNITIF	rosārum	-ārum
	DATIF	rosis	-is
	ABLATIF	rosis	-is

PARTICULARITÉS DE LA 1^{re} DÉCLINAISON

13. **Família**, ae, f. : famille a gardé un génitif sg. archaïque en -as dans paterfamília, patrisfamília, m. materfamília, matrisfamília, f. père de famille mère de famille
14. **Dea**, deae, f. : déesse et **filia**, ae, f. : fille font au datif et à l'ablatif pluriels **deabus** et **filiabus** quand il faut éviter une confusion avec le dat.-abl. pl. de deus, dei, m. : dieu et filius, ii, m. : fils.

Deis et **deabus** : Aux dieux et aux déesses.
Filiis et **filiabus** : Aux fils et aux filles.

→ Mais on peut trouver **deis**, **filiis** venant de **dea**, **filia**, s'il n'y a pas de risque de confusion.

15. Formes poétiques : gén. sg. en -āi au lieu de -ae : terrāi
gén. pl. en -um au lieu de -ārum : agricolum.

This talk

1. Why am I interested in data management ?
2. Current problems
3. Data principles and concrete step to apply them
4. The Paralex standard for inflected lexicons

Why?

Variability in paradigm sizes

Vietnamese: 1 form

(1) Tôi học
I study
'I study'

(2) Tôi đang học
I FUT study
'I will study'

(3) Tôi đã học
I PST study
'I studied'

English: 5 forms

STUDY

PRS stʌɪ

PRS.3SG stʌɪz

PST stʌɪd

PST.PART stʌɪdɪŋ

PRS.PART stʌɪɪŋ

French: 50 forms

INDICATIF	
<ul style="list-style-type: none"> • PRÉSENT <ul style="list-style-type: none"> j étudie tu étudies il étudie elle étudie nous étudions vous étudiez ils/elles étudient • IMPARFAIT <ul style="list-style-type: none"> j étudiais tu étudiais il étudiait elle étudiait nous étudions vous étudiez ils/elles étudiaient • PASSÉ SIMPLE <ul style="list-style-type: none"> j étudiai tu étudias il étudia elle étudia nous étudîmes vous étudîtes ils/elles étudient • FUTUR SIMPLE <ul style="list-style-type: none"> j étudierai tu étudieras il étudiera elle étudiera nous étudierons vous étudierez ils/elles étudieront 	<ul style="list-style-type: none"> • PASSÉ COMPOSÉ <ul style="list-style-type: none"> j ai étudié tu as étudié il a étudié elle a étudié nous avons étudié vous avez étudié ils/elles ont étudié • PLUS QUE PARFAIT <ul style="list-style-type: none"> j avais étudié tu avais étudié il avait étudié elle avait étudié nous avions étudié vous aviez étudié ils/elles avaient étudié • PASSÉ ANTERIEUR <ul style="list-style-type: none"> j eus étudié tu eus étudié il eut étudié elle eut étudié nous eûmes étudié vous eûtes étudié ils/elles eussent étudié • FUTUR ANTERIEUR <ul style="list-style-type: none"> j aurai étudié tu auras étudié il aura étudié elle aura étudié nous aurons étudié vous aurez étudié ils/elles auront étudié
CONDITIONNEL	
<ul style="list-style-type: none"> • PRÉSENT <ul style="list-style-type: none"> j étudierais tu étudierais il étudierait elle étudierait nous étudierions vous étudieriez ils/elles étudieraient 	<ul style="list-style-type: none"> • PASSÉ <ul style="list-style-type: none"> j aurais étudié tu aurais étudié il aurait étudié elle aurait étudié nous aurions étudié vous auriez étudié ils/elles auraient étudié
SUBJONCTIF	
<ul style="list-style-type: none"> • PRÉSENT <ul style="list-style-type: none"> que j étudie que tu étudies qu'il étudie qu'elle étudie que nous étudions que vous étudiez qu'ils/elles étudient • IMPARFAIT <ul style="list-style-type: none"> que j étudiasse que tu étudias qu'il étudiat qu'elle étudiat que nous étudissions que vous étudissiez qu'ils/elles étudissent 	<ul style="list-style-type: none"> • PASSÉ <ul style="list-style-type: none"> que j aie étudié que tu aies étudié qu'il ait étudié qu'elle ait étudié que nous ayons étudié que vous ayez étudié qu'ils/elles aient étudié • PLUS QUE PARFAIT <ul style="list-style-type: none"> que j eusse étudié que tu eusses étudié qu'il eût étudié qu'elle eût étudié que nous eussions étudié que vous eussiez étudié qu'ils/elles eussent étudié
IMPERATIF	
<ul style="list-style-type: none"> • PRÉSENT 	<ul style="list-style-type: none"> • PASSÉ

Archi: 1 500 000 forms (Dagestan)

Variability in paradigm structures

English

cat 'one cat'
cat-s 'many cats'

Yei (New Guinea)

yə-mdəd-ə 'one sat'
yε-mdəd-anε 'two sat'
yε-mdəd-ə 'many sat'

Walpiri (Australia)

wati 'one man'
wati-jarra 'two men'
wati-patu 'several men'

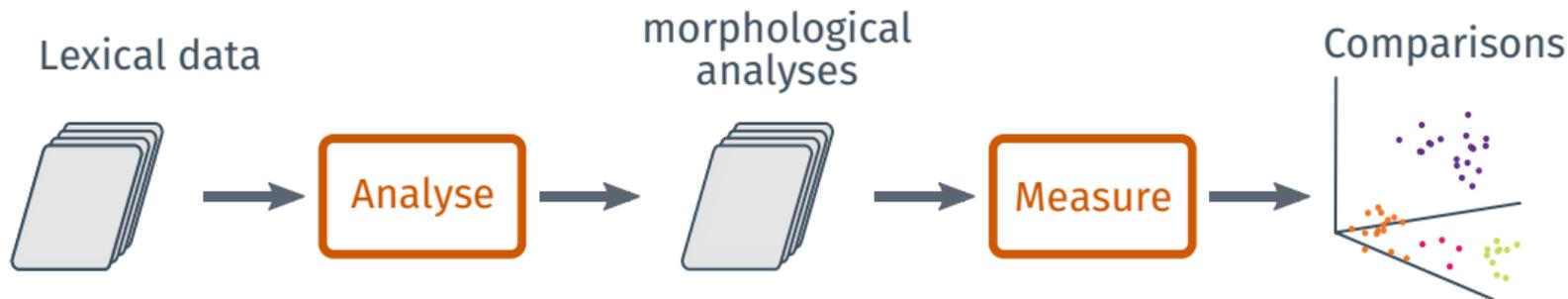
Nuer (Soudan)

t̪ùrbîl 'one cat'
t̪ùrbîεel 'of a cat'
t̪ùɔrbĩεelĩ 'many cats'

Research questions

- (1)** *What is possible?*
characterize the variability space
- (2)** *What is observed ?*
Find macro-linguistic variation distributions
- (3)** *Why?*
Testable hypotheses on cognition, cultural evolution, communication

Computational methodology



Large scale & high resolution multivariate typology:

- Precise, high quality data;
- Automated linguistic analyses
- Quantitative measurements

Why not the unimorph datasets ?

Kirov et al. (2016) and Batsuren et al. (2022):

```
haarama olete haaranud V;ACT;PRS;PRF;POS;IND;2;PL
haarama ei olevat haaranud V;ACT;PRS;PRF;NEG;QUOT
haarama oleks haaratud V;PASS;PRS;PRF;POS;COND
haarama haarasin V;ACT;PST;POS;IND;1;SG
haarama haاراتakse V;PASS;PRS;POS;IND
haarama haaranud V.PTCP;ACT;PST
haarama haاراتagu V;PASS;PRS;POS;IMP
haarama արցս haարակս V;ACT;PRS;NEG;IMP;3;SG
haarama oli haاراتս V;PASS;PST;PRF;POS;IND
haarama haarasime V;ACT;PST;POS;IND;1;PL
haarama ei olnud haaranud V;ACT;PST;PRF;NEG;IND
haarama oleksid haaranud V;ACT;PRS;PRF;POS;COND;2;SG
haarama olgu haaranud V;ACT;PRS;PRF;POS;IMP;PL
haarama oli haaranud V;ACT;PRS;PRF;POS;COND;3;PL
haarama haاراتaks V;PASS;PRS;POS;COND
haarama haարaksime V;ACT;PRS;POS;COND;1;PL
haarama on haاراتս V;PASS;PRS;PRF;POS;IND
haarama արցս haարակս V;ACT;PRS;NEG;IMP;2;PL
haarama haարab V;ACT;PRS;POS;IND;3;SG
haarama haարavat V;ACT;PRS;POS;QUOT
haarama արցս haարակս V;ACT;PRS;NEG;IMP;3;PL
haarama olgu haaranud V;ACT;PRS;PRF;POS;IMP;SG
haarama haարատի V;PASS;PST;POS;IND
haarama ei haարavat V;ACT;PRS;NEG;QUOT
haarama ei oleks haاراتս V;PASS;PRS;PRF;NEG;COND
haarama ei haaranud V;ACT;PST;NEG;IND
haarama ei olevat haاراتս V;PASS;PRS;PRF;NEG;QUOT
haarama ei haարaks V;ACT;PRS;PRF;POS;COND;1;SG
haarama haاراتս V.PTCP;PASS;PST
haarama olid haaranud V;ACT;PRS;PRF;POS;COND;2;SG
haarama haարavad V;ACT;PRS;POS;IND;3;PL
haarama olite haaranud V;ACT;PRS;PRF;POS;COND;2;PL
haarama ei haاراتս V;PASS;PST;NEG;IND
haarama oleksite haaranud V;ACT;PRS;PRF;POS;COND;2;PL
haarama haարav V.PTCP;ACT;PRS
haarama haարակս V;ACT;PRS;POS;IMP;3;SG
haarama haarama V;NFIN
```

- Crowdsourced origin
- Minimalist, **canonical** structure
- Strictly **orthographic**
- Large but **noisy**

<https://unimorph.github.io/>

Problems

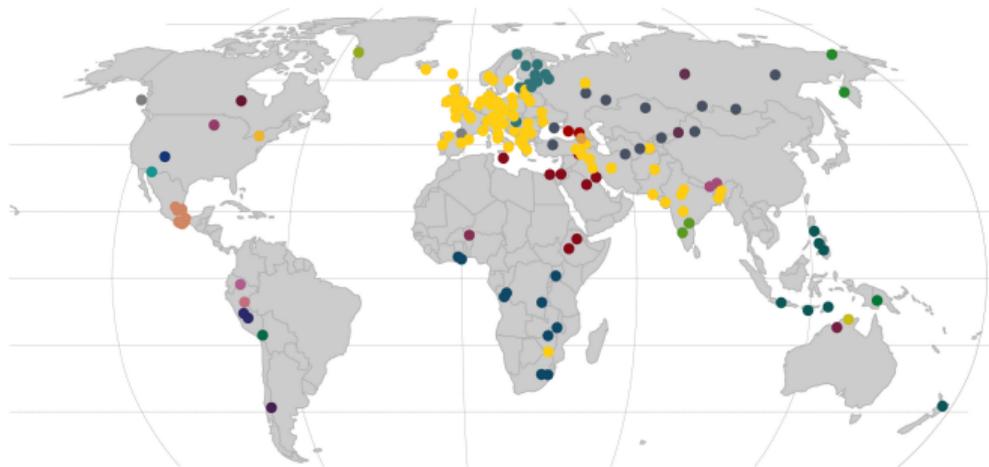
Culture

“You can also collect butterflies and make many observations. If you like butterflies, that’s fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so.”

Chomsky and Ronat, 1979, p. 57



Coverage

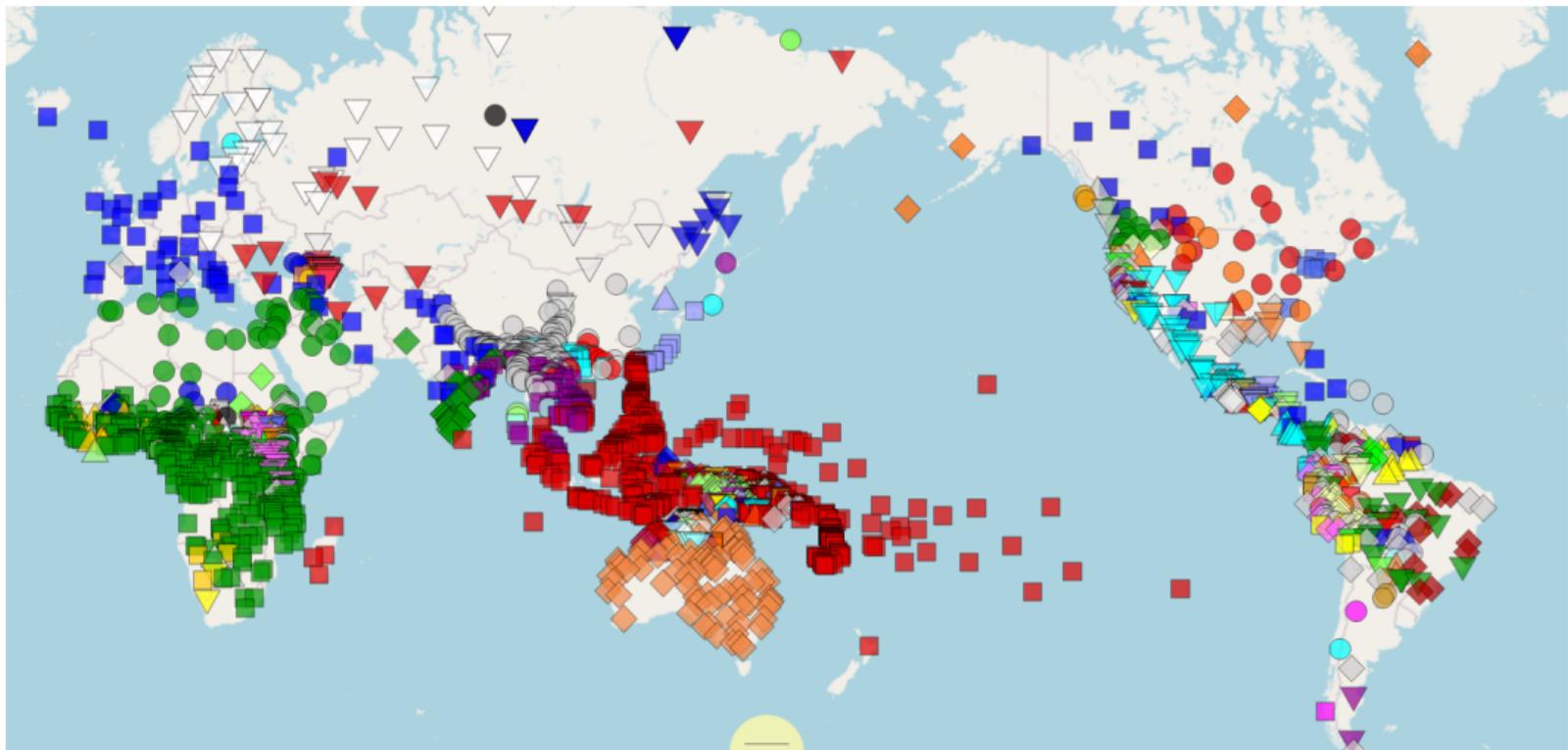


Macro-family

- | | | | | |
|---------------------------|------------------|----------------|-----------------|-----------------|
| ● Abkhaz-Adyge | ● Afro-Asiatic | ● Algic | ● Araucanian | ● Arawakan |
| ● Athabaskan-Eyak-Tlingit | ● Atlantic-Congo | ● Austronesian | ● Aymaran | ● Bosavi |
| ● Chukotko-Kamchatkan | ● Dravidian | ● Eskimo-Aleut | ● Gunwinyguan | ● Indo-European |
| ● Iroquoian | ● Kartvelian | ● Otomanguean | ● Pano-Tacanan | ● Quechuan |
| ● Sino-Tibetan | ● Siouan | ● Songhay | ● Southern Daly | ● Tungusic |
| ● Turkic | ● Uralic | ● Uto-Aztecan | ● NA | |

Coverage

Grambank (Skirgård et al., 2023): <https://grambank.clld.org/>



Consistency and machine-readability

Gap between human- and machine-readability:

- **Formats**
- **Coding**
- Down to details of encoding

Commensurability

- Linguistic data are analyses
- Information of absence vs absence of information
- Interpretation of coding schemes

Durability

Romance Verbal Inflection Dataset 2.0 Beniamine et al (2020)

Data imprisoned within its own website

web scraping, cleaning,
normalisation, standardisation,
publication

<https://doi.org/10.5281/zenodo.4039059>

Surrey Morphological Complexity Database

Data initially published in Flash,
unsupported since 2021.

Data archeology, cleaning,
updating, normalisation,
publication

<http://dx.doi.org/10.15126/SMG.23/1>

Technical skills

Rarely do we teach linguistics students about:

- Machine readable formats (xml, json, csv)
- Versioning (git)
- Data cleaning, validation, testing
- Metadata, standards

Some solutions

Data principles



Some Paralex collaborators

Cormac Anderson



Jules Bouton



Mae Carroll



Borja Herce



Matteo Pellegrini



Erich Round



Helen Sims-Williams



DeAR Principles

- **Decentralized**
 - Centralized data is neither robust nor long-lasting
 - International collaboration must be incentivised
- **Automatically validated**
 - Manual curation of large datasets is necessary but error-prone
 - Automatic quality control of structure and content is crucial
- **Revisable**
 - Data evolves, its presentation must be updateable
 - Seamless updates of showcases can be generated from a single data source

Beniamine et al 2023. *Paralex: a DeAR standard for rich lexicons of inflected forms*. ISMo: International Symposium of Morphology.

Solutions

To publish high quality, easily citable, scientifically impactful data, useful for the long term:

Creation

- Versioning **DeAR**
- Metadata **FAIR** **DeAR**
- Standards **Inter-operable** **Reusable** **DeAR**
- Linked data **Interoperable** **Reusable**
- Validation **DeAR**

Publication

- Documentation **Reusable**
- Licenses **Open** **Reusable**
- Archived downloads **FAIR**
- DOIs **Findable** **Accessible**
- Continuous pipelines **DeAR**

Standards

For data points

- Metric systems
- Leipzig glossing rules
- ISO 639 (langues)
- ISO 3166 (pays)

For datasets

- Text Encoding Initiative (TEI, XML)
- CLDF
- CONLL-U
- Paralex !

For metadata

- Frictionless (json)
- Dublin Core (xml)

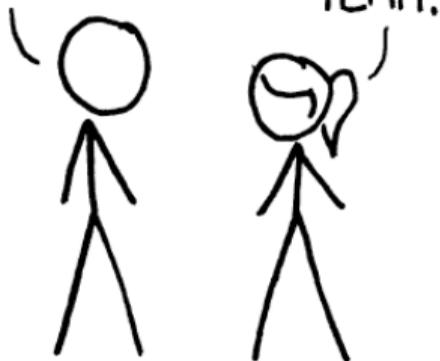
The Paralex standard

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.



SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

[Paralex](#)[Home](#)[Standard](#)[Specs](#)[Ontology](#)[Background](#)[Tutorial](#)[Datasets](#)[Changelog](#)

Paralex: lexicons of morphological paradigms

pypi package 2.2.1

Paralex is a standard for morphological lexicons which document inflectional paradigms.

It strives to provide data which is [FAIR](#), so it can be used automatically, [CARE](#), so it respects and empowers language communities, and [DeAR](#) (our own set of principles), so we can create a virtuous data ecosystem. It was inspired by the [Cross-Linguistic Data Formats \(CLDF\)](#) standard, and adheres to a similar philosophy. We aim to keep the two standards compatible in order to facilitate inter-operability.

LINKED

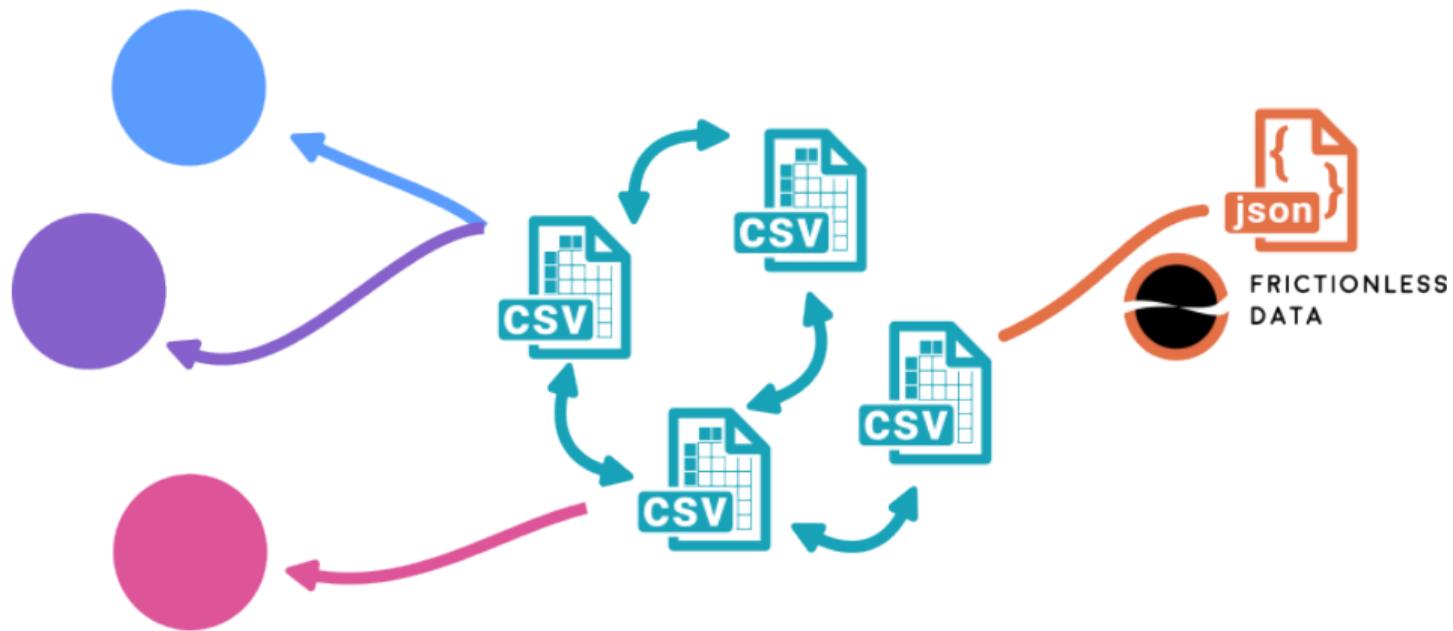
DATA

METADATA

LINKED

DATA

METADATA



Relational schema

form_id	lexeme	cell	phon_form	orth_form
f1	CHANTER	PRS.IND.1.SG	ʃã t	chante
f4	CHANTER	PRS.IND.1.PL	ʃã t õ	chantons
f60	PELER	PRS.IND.1.SG	p ε l	pèle
f64	PELER	PRS.IND.1.PL	p ø l õ	pelons
f90	FINIR	PRS.IND.1.SG	f i n i	finis
f94	FINIR	PRS.IND.1.PL	f i n i s õ	finis

forms	
form_id	string
lexeme	string
cell	string
phon_form	string
orth_form	string

Relational schema

form_id	lexeme	cell	phon_form	orth_form
f1	CHANTER	PRS.IND.1.SG	ʃã t	chante
f4	CHANTER	PRS.IND.1.PL	ʃã t õ	chantons
f60	PELER	PRS.IND.1.SG	p ε l	pèle
f64	PELER	PRS.IND.1.PL	p ø l õ	pelons
f90	FINIR	PRS.IND.1.SG	f i n i	finis
f94	FINIR	PRS.IND.1.PL	f i n i s õ	finis

lexeme_id	inflection_class	gloss
CHANTER	groupe-1	to eat
PELER	groupe-1	to peel
FINIR	groupe-2	to end

forms	
form_id ↻	string
lexeme	string
cell	string
phon_form	string
orth_form	string

lexemes	
lexeme_id ↻	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

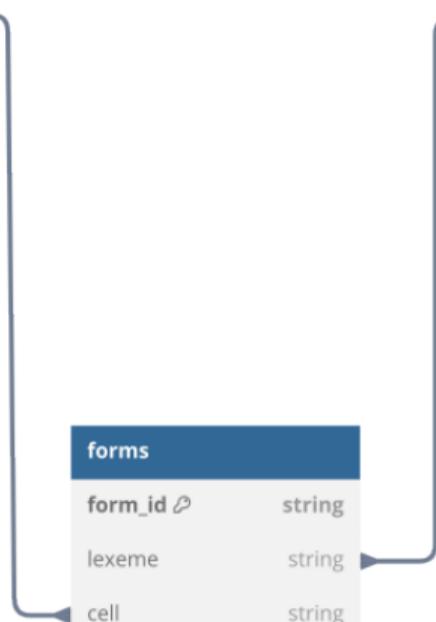
Relational schema

cells	
cell_id ↗	string
POS	string
unimorph	string
ud	string
comment	string

lexemes	
lexeme_id ↗	string
inflection_class	string
meaning	string
gloss	string
POS	string
comment	string

cell_id	unimorph	POS
IND.PRS.1.SG	V;IND;PRS;1;SG	verb
IND.PRS.1.PL	V;IND;PRS;1;PL	verb

forms	
form_id ↗	string
lexeme	string
cell	string
phon_form	string
orth_form	string



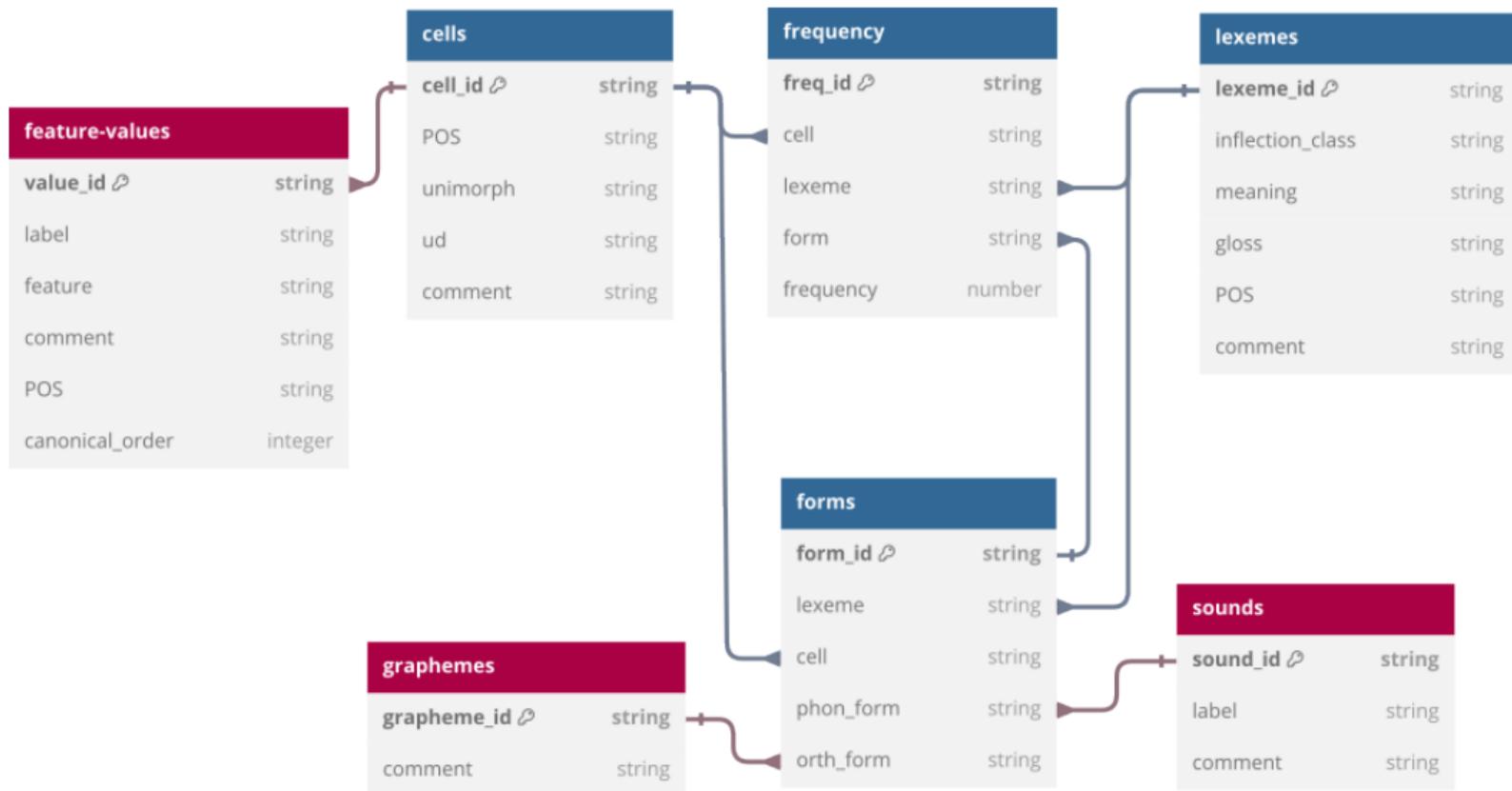
Relational schema

cell_id	GRACE	flexique	unimorph	ud	ftb
COND.PRS.1.PL	Vmcp1p-	cond.1pl	V;COND;1;PL	Mood=Cnd Number=Plur Person=1 Tense=Pres	m=cond n=p p=1 t=pst
COND.PRS.1.SG	Vmcp1s-	cond.1sg	V;COND;1;SG	Mood=Cnd Number=Sing Person=1 Tense=Pres	m=cond n=s p=1 t=pst
COND.PRS.2.PL	Vmcp2p-	cond.2pl	V;COND;2;PL	Mood=Cnd Number=Plur Person=2 Tense=Pres	m=cond n=p p=2 t=pst
COND.PRS.2.SG	Vmcp2s-	cond.2sg	V;COND;2;SG	Mood=Cnd Number=Sing Person=2 Tense=Pres	m=cond n=s p=2 t=pst
COND.PRS.3.PL	Vmcp3p-	cond.3pl	V;COND;3;PL	Mood=Cnd Number=Plur Person=3 Tense=Pres	m=cond n=p p=3 t=pst
COND.PRS.3.SG	Vmcp3s-	cond.3sg	V;COND;3;SG	Mood=Cnd Number=Sing Person=3 Tense=Pres	m=cond n=s p=3 t=pst
IND.FUT.1.PL	Vmif1p-	fut.1pl	V;IND;FUT;1;PL	Mood=Ind Number=Plur Person=1 Tense=Fut	m=ind n=p p=1 t=fut
IND.FUT.1.SG	Vmif1s-	fut.1sg	V;IND;FUT;1;SG	Mood=Ind Number=Sing Person=1 Tense=Fut	m=ind n=s p=1 t=fut
IND.FUT.2.PL	Vmif2p-	fut.2pl	V;IND;FUT;2;PL	Mood=Ind Number=Plur Person=2 Tense=Fut	m=ind n=p p=2 t=fut
IND.FUT.2.SG	Vmif2s-	fut.2sg	V;IND;FUT;2;SG	Mood=Ind Number=Sing Person=2 Tense=Fut	m=ind n=s p=2 t=fut
IND.FUT.3.PL	Vmif3p-	fut.3pl	V;IND;FUT;3;PL	Mood=Ind Number=Plur Person=3 Tense=Fut	m=ind n=p p=3 t=fut
IND.FUT.3.SG	Vmif3s-	fut.3sg	V;IND;FUT;3;SG	Mood=Ind Number=Sing Person=3 Tense=Fut	m=ind n=s p=3 t=fut
IMP.PRS.1.PL	Vmmp1p-	imp.1pl	V;POS;IMP;1;PL	Mood=Imp Number=Plur Person=1 Tense=Pres	m=Imp n=p p=1 t=pst
IMP.PRS.2.PL	Vmmp2p-	imp.2pl	V;POS;IMP;2;PL	Mood=Imp Number=Plur Person=2 Tense=Pres	m=Imp n=p p=2 t=pst
IMP.PRS.2.SG	Vmmp2s-	imp.2sg	V;POS;IMP;2;SG	Mood=Imp Number=Sing Person=2 Tense=Pres	m=Imp n=s p=2 t=pst
INF	Vmn--	inf	V;NFIN	VerbForm=Inf	m=inf
IND.IPFV.1.PL	Vmii1p-	ipfv.1pl	V;IND;PST;1;PL;IPFV	Mood=Ind Number=Plur Person=1 Tense=Imp	m=ind n=p p=1 t=Imp
IND.IPFV.1.SG	Vmii1s-	ipfv.1sg	V;IND;PST;1;SG;IPFV	Mood=Ind Number=Sing Person=1 Tense=Imp	m=ind n=s p=1 t=Imp
IND.IPFV.2.PL	Vmii2p-	ipfv.2pl	V;IND;PST;2;PL;IPFV	Mood=Ind Number=Plur Person=2 Tense=Imp	m=ind n=p p=2 t=Imp
IND.IPFV.2.SG	Vmii2s-	ipfv.2sg	V;IND;PST;2;SG;IPFV	Mood=Ind Number=Sing Person=2 Tense=Imp	m=ind n=s p=2 t=Imp
...

Vocabulary relations

form_id	lexeme	cell	phon_form
form_158742	mésuser	ptcp.pst.f.pl	m E z y z e
form_819	aboyer	ind.prs.1.pl	a b w a j ã
form_41745	chroniquer	cond.prs.1.sg	k ʋ O n i k ə ʋ E
form_91334	détricoter	ind.pst.3.pl	d E t ʋ i k O t E ʋ
form_197935	refleurir	imp.prs.2.sg	ʋ ə f l Ø ʋ i
form_122951	galvaniser	ind.pst.3.sg	g a l v a n i z a
form_11785	anoblir	ind.prs.2.pl	a n O b l i s e
form_99328	encourager	ind.prs.2.pl	ã k u ʋ a ʒ e
form_237143	surprendre	sbjv.prs.1.pl	s y ʋ p ʋ ə n j ã
...

Vocabulary relations



Frequencies

We need frequencies to go beyond idealized paradigm tables:

- Effect on paradigm structure
- Studying non-canonical phenomena

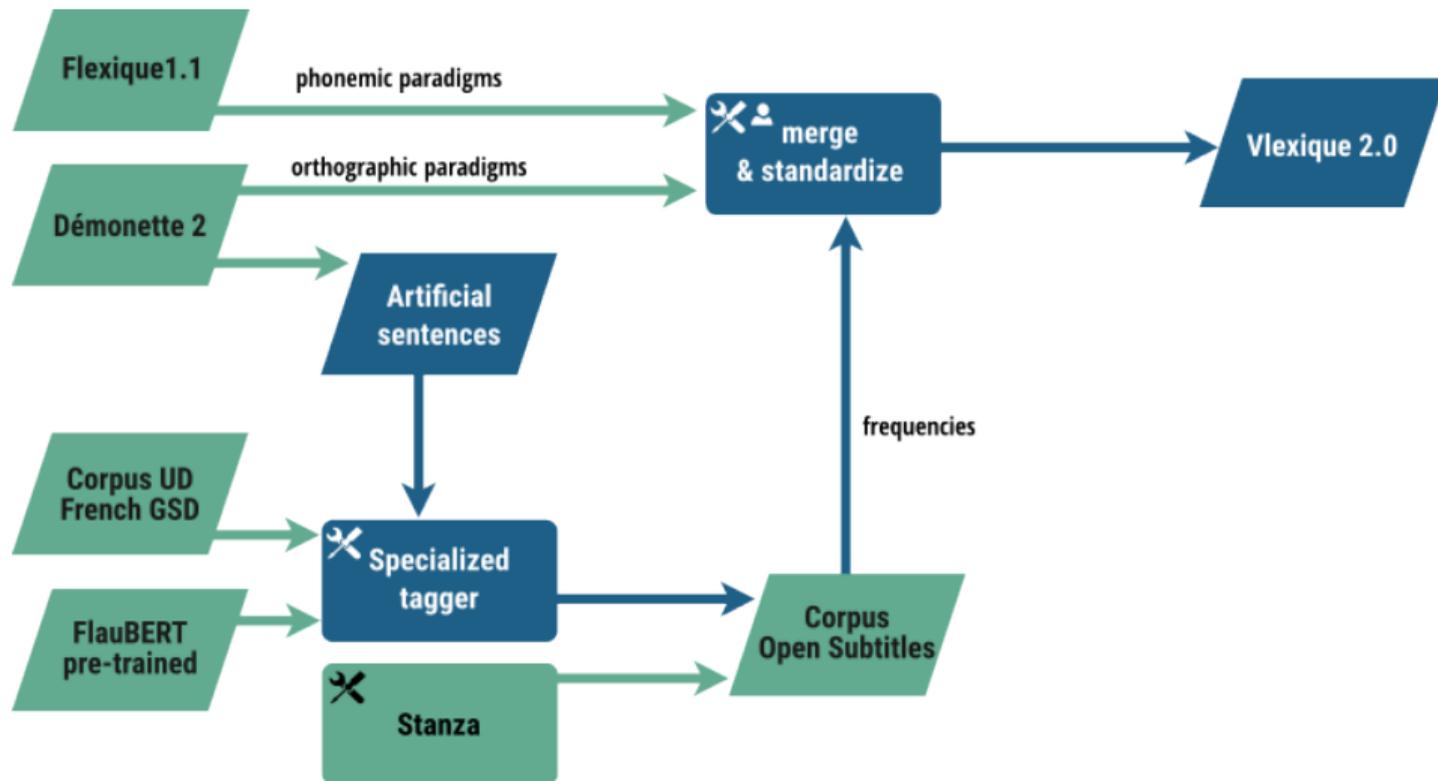
But obtaining good frequencies is hard:

- A** Skewed frequencies in annotated corpora
- B** Syncretic forms within lexemes
- C** Homonymous forms across lexemes

Orthographic tokens and inflected forms

- (4) a. Elle savait que vous **veniez**
she.3SG.F knew.IMP.3SG COMP you.2PL come.**IND.IPFV.2PL**
'She knew that you were coming.'
- b. Il faudrait que vous **veniez**
one.3SG.M must.COND.3SG COMP you.2PL come.**SBJV.PRS.2PL**
'You should come.'
- (5) a. Je **comparais** devant une cour [...] (**COMPARAÎTRE**)
I.1SG appear.IND.PRS.1.SG before a court
'I appear before a court'
- b. Je **comparais** juste les tailles ! (**COMPARER**)
I.1SG compare.IND.IPFV.1.SG only the sizes !
'I was only comparing sizes' !

Vlexique creation



Beniamine, Coavoux & Bonami (2024). *Vlexique2.0: A rich lexicon of French verbal inflection with form-level frequencies*. 21st International Morphology Meeting, Aug 2024, Vienna, Austria. hal-04689352

Generated websites

ARAVel.ex: Modern Standard Arabic Verbal lexicon

Home About Cite Forms Lexemes Search Features release

This work is licensed under [CC BY-NC 4.0](#)

See the [automatically generated cite for these paradigms](#)

Aravel.ex: Modern Standard Arabic Verbal lexicon

Cite as: Beniamine, Sacha (2021). *Aravel.ex: Modern Standard Arabic Verbal Lexicon* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5121343>

This is an updated version of the lexicon published as part of *Generative Grammar 2019, July*, Classification: ResearchGate. Under quantitative data collection de paradigms. Ph. D. thesis, Université Sorbonne Paris Cité / Université Paris Diderot

Which itself is based on the 2016 version of the *Leximorph* dataset for Arabic. We kept that original file in `2016`, as the recent versions found at <https://github.com/leximorph/leximorph-master> tend to not present as much information (in particular, they do not include the transcription). See [Klein, Christy et al. \(2019\)](https://arxiv.org/abs/1907.02489), "Very large scale parsing and normalization of written Arabic Morphological Paradigms", in: *Proceedings of the 16th International Conference on Language Resources and Evaluation (LREC 2019)*, Shikha Al-Ahmed, Nicola Colliard, Christoforos Christal et al., Portland, Oregon: International Language Resources Association (LREC), 4793-4799. <https://arxiv.org/abs/1907.02489>

Phonemic transcriptions are obtained automatically using rules, starting from the [orthographic transcription](#), which is [independently prepared](#), and validated by a native speaker. A number of changes were made with respect to the dataset from Beniamine (2018):

- The dataset is fully Paralex compliant, and includes tables for calls, lexemes, bases and roots
- Transcription changes:
 - We use `ʔ` rather than `g` for pharyngeal consonants, with a suitable set of rules (see `2018/19/1`)
 - The letter `ð` (`ṯ`) is always transcribed `ʃ`, even for `ʃ` when it is a number of user entries is followed by evidence comments (Beniamine (2018) entry 4)
 - When there was a single consonant for multiple orthographic forms, we keep only the first one
 - Calls and features changes:
- We use the terms `imperfect (paraf)` and `perfect (parf)` rather than `present indicatif` and `future indicatif`
- Mapping for calls is given from our system to both the old `araphoncode` (in `2016`) and the new one (found at <https://github.com/leximorph/leximorph-master>)
- Lexemes change (different verbal paradigms from the same page have been distinguished)

Re-generating data

European Portuguese Verbs

Home About Cite Forms Lexemes Search Features release

Home

DOI: 10.5281/zenodo.5121343

This is a collection of European Portuguese verbal paradigms, in phonemic notation. They are suited for both computational and manual analysis. The `paradigms` table lists all available lexemes, and provides full paradigms for each. The `signposts` table lists all phonemes used in the transcription, and describes them in terms of distinctive features.

The *European Portuguese Verbs lexicon* is licensed under [Attribution-ShareAlike 4.0 International](#)

Please cite as:

- Perdigão, Fernando, Beniamine, Sacha, Luís, Ana R., & Bonami, Olivier. (2021). *European Portuguese Verbal Paradigms in Phonemic Notation* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5121343>

Version 1.0.1 of this lexicon was prepared for the publication:

- Sacha Beniamine, Olivier Bonami, and Ana R. Luís (2021). The fine implicative structure of European Portuguese conjugation. *Isoglossa*. *Open Journal of Romance Linguistics*. DOI: <https://doi.org/10.5565/rev/isoglossa.109>

The data can be downloaded from [Zenodo](#) or from the [github repository](#).

How this lexicon was prepared

Esthetic: A Paralex Lexicon of Estonian Paradigms

Home About Cite Forms Lexemes Search Features release

DOI: 10.5281/zenodo.5121343

Esthetic is a collection of Estonian verbal and nominal paradigms, in phonemic and orthographic notation. They are suited for both computational and manual analysis.

The data files are encoded in `csv` files, and the metadata follows [Fric/Slocus](#) standards. The dataset conforms to the [Paralex standard](#)

The *Estonian Paradigms in Phonemic Notation* dataset is licensed under [Attribution-ShareAlike 4.0 International](#)

Please cite as:

- Sacha Beniamine, Mari Algo, Matthew Baerman, Jules Bouton & Marie Copot (2024). *Esthetic: A Paralex Lexicon of Estonian Paradigms*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. To appear.
- Sacha Beniamine, Mari Algo, Matthew Baerman, Jules Bouton & Marie Copot. *Estonian Paradigms in Phonemic Notation* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8383523>

The data can be downloaded from [Zenodo](#) or from the [github repository](#), and [visualized online](#).

We thank Indrek Hein and Ylle Viik for providing us access to Ekilex data and plenty of support.

References

This dataset is derived from Ekilex data. See:

- Ekilex API <https://github.com/tripledex/ekilex/wiki/Ekilex-API>

Ngkolmpu verbal paradigms

Ngkolmpu Verbal Paradigms Home About Cite Forms Lexemes Search

Ngkolmpu Verbal Paradigms

Ngkolmpu Paralex Verbal Lexicon

Table of contents
About
How this lexicon was prepared
The data
Contact
References

About

This is a collection of selected forms of verbs in Ngkolmpu, a Creole language spoken in the Indonesian Province of Papua Selatan in the southern New Guinea region.

The paradigms are currently presented in phonemic notation and closely follow the analysis presented in Carstairs (2016).

The data is encoded in `csv` files, and the metadata follows [Fric/Slocus](#) standards. The dataset conforms to the [Paralex standard](#).

How this lexicon was prepared

All the data here was collected in consultation with Ngkolmpu speakers in the village of Yenggorobu, Merauke Regency, South Papua Province, Indonesia.

The data

There are currently 104 verb lexemes in the collection. This does not include derived or alternations for `unstressed`.

The current list contains a highly phonemized version which doesn't represent typologically conditioned alternations.

This is a work in progress.

Contact

Any questions please contact Ilse Carstairs email=carstairs@wfu.edu

References

Carstairs, M.L. 2016. The Ngkolmpu Language with special-interest in distributed exponence.

Documentation and data sheet

 **Vlexique**

Docs Lexemes Forms Cells Sounds Tags Features values

Docs
Home
Transcription
[Data Sheet](#)
Metadata
References

Data set name: Vlexique: French Verbal Paradigms in Phonemic Notation

Citation (if available): Beniamine, Coavoux, Bonami. Vlexique2.0 [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.10638682>

Data set developer(s): Sacha Beniamine

Data sheet author(s): Sacha Beniamine

Others who contributed to this document: None

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created in order to provide data on French Verbal Inflection. It is intended for use in NLP and linguistic investigation.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

This dataset was created by Sacha Beniamine¹, Maximin Coavoux², and Olivier Bonami³. It is largely based on [Flexique v.1.3](#); itself derived from [Lexique](#). Orthographic forms are taken from [Démonette](#).

Affiliations: ¹Surrey Morphology Group, University of Surrey, United Kingdom ²Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France ³Laboratoire de Linguistique Formelle, Université Paris Cité, France

- Document analytical choices
- Document any hesitations or decisions in coding
- Data sheet: targeted questions mixing data statements (Bender and Friedman, 2018) & data sheets (Gebru et al., 2021)

Current coverage



■ Yam (more1255)

■ Otomanguean (otom1299)

■ Pama-Nyungan (pama1250)

■ Atlantic-Congo (atla1278)

■ Uralic (ural1272)

■ Sino-Tibetan (sino1245)

■ Indo-European (indo1319)

■ Afro-Asiatic (afro1255)

Conclusion

- Computational linguistics benefits from datasets which are:
 - high quality
 - long lasting
 - open, FAIR, CARE, & DeAR
- Large noisy datasets are not enough
- A crucial point is to lower the technical bar
- Good data are among the most important contributions we can make to linguistics

Conclusion

- Computational linguistics benefits from datasets which are:
 - high quality
 - long lasting
 - open, FAIR, CARE, & DeAR
- Large noisy datasets are not enough
- A crucial point is to lower the technical bar
- Good data are among the most important contributions we can make to linguistics

Thanks !